**ISO/IEC JTC 1/SC 2/WG 2 *N2661***

***Date: 2003-10-17***

# ISO/IEC JTC 1/SC 2/WG 2

**Universal Multiple-Octet Coded Character Set (UCS) - ISO/IEC 10646**

**Secretariat: ANSI**

| | |
|---|---|
| **Title:** | **Clarification and Explanation on Tibetan BrdaRten Proposal** |
| **Precedes:** | WG2 N2558, N2621 and Critique Papers N2624, N2635, N2637 and N2638 |
| **Source:** | **China National Body** |
| **Action:** | **To Discuss at WG2 #44** |
| **Distribution:** | **SC2/WG2 Experts** |

This paper is to further clarify and explain issues in N2621, the Revised Proposal on Tibetan BrdaRten, and reply the critiques in Papers N2624, N2635, N2637 and N2638.

The Chinese National Body would like to reaffirm its position , that the encoding Model, Tibetan Basic Letter Plus BrdaRten has been being used in Tibetan language for machine processing, which is the most suitable model for China to facilitate and speed-up the migration of Tibetan system and e-data in China to UCS, therefore, China insists on the proposal to encode BrdaRten in to BMP of UCS for interoperability.

## 1. Why BrdaRten? Is BrdaRten necessary ?

China does not intend to challenge WG2 general principle on pre-composed character encoding, but would like to objectively analyze the specific language/script for its encoding.

(1) **Tibetan BrdaRten has been being an element for machine processing** since typewriter age, which is ease to recognize by eyes, ease to process by machine. Only because of the capacity of the machine, the number of used BradaRten is limited to small amount. The lead letters and keyboard of the Tibetan typewriter are shown in Figure 1 and 2.

***Figure 1. Lead Lettersof BrdaRten***

***Figure 2: Typewriter with BrdaRten Keyboard.***


(2) When PC and PC based desk-top typesetting system is used in Tibetan language area, the BrdaRten characters are adopted in large scale for input, storage, processing and display-printing in China. According to the statistics based on the Tibetan data base by The NorthWest Minority Nationality University, among the 18 millions character text, the basic Tibetan letters appear 40.97%, while the BrdaRten is **23.07%** (the Intersyllabic Mark 0F0B and Phrase Separator 0F0D occupy 34.73% ). Another statistics jointly made by China Institute of Standardization and Qinghai Normal University is similar :45.62% and **23.59%** based on 19 millions character data base. **This implies that every two basic Tibetan letters accompany one more BrdaRten in average** ! This high frequency hints us that the BrdaRten form could not be ignored.

(3) Since early 1990s, the Tibetan systems by Founder , Huaguang and Tongyuan in later, are widely used in newspapers, magazines, publishing Houses, Central and Local Translation Bureaus, Tibetology Research Center, local governments , Tibetan Radio /TV Stations in Gansu, Qinghai, Sichuan, Yunnan and Tibet. All of official publications are utilizing Tibetan basic letters plus BrdaRten coding model although the code pages amongst Founder, Huaguang and Tongyuan are different in repertoire and code position. It is roughly estimated that **such users exceed 3000** (for registered users, Founder – 600-700, Huaguang – 700-800, Tongyuan – 1500, BanZhiDa- >100 ), the e-data with BradaRten exceeds **several billions characters**. This huge amount of e-data with mission-critical users is another fact which can not be ignored.

Therefore, encoding BrdaRten as it is in UCS/BMP is necessary because of its history, its usage practice , its usage frequency, and the huge amount electronic data existed (legacy) in the coding model. The next paragraph will give more fact and reasons compared with the dynamically-combining model.

## 2. Is BrdaRten Sufficient ?

The answer is **Yes in general**. The proposed BrdaRten characters in N2621 are **sufficient for modern Tibetan language**. Together with Tibetan basic characters, BrdaRten can cover 100% modern Tibetan written language, even most ancient Tibetan classics. This is the goal of N2621.

On the other hand, in order to meet the requirement for digital Tibetan ancient classics, **another 6000 BrdaRten character set** has been prepared for national standard pending on its code position, which may be placed in Private Use Zone, or in Supplementary Plane. Actually, such rarely used BrdaRten characters in Huaguang system are being used for publishing and proof-reading the large-scale classics

Tibetan Buddhism 藏文《大藏经 甘珠尔 丹珠尔》(*bkav bstan vgyur, bstan vgyur and bkav bstan*), which are printed in 290,000 pages, about 290 millions Tibetan characters containing BrdaRten.

Nevertheless, proposed and to be proposed BrdaRten set would not be able to exhaustedly cover any STACKS , just like N2624 said, there is also *exception even for existing Unicode character encoding model, t*he *highly unusual stacks that contain more than one consonant-vowel combination in a vertical arrangement (these contravene the normal rules of Tibetan writing, and are considered beyond the scope of plain text rendering).*

## 3. BrdaRten vs. Dynamically-Combining, Pros and Cons

Anything has its positive side and negative side, so does BdaRten Model or Combining Model for Tibetan. What Chinese Tibetan people need is an advanced and mature technique to implement the model easily, quickly and at lower cost.

We carefully studied and tested the beta version Uniscribe Complex-Script Rendering Engine with Ximalaya OpenType Font as an implementation of Dynamically-Combining Model purely based on 0Fxx . It looks smart and interesting, but still **in developing stage**. It would be a good solution in the future, but, unfortunately, it has **not proved as a MATURE product yet**, in user-friendliness for BrdaRten key-in and editing.

Rather than arguing whether or not this is due to the coding scheme, we would like to make a comparison of dynamically-combining model with BrdaRten model. For brevity, 0FXX is used for Dynamically-Combining Coding Model for Tibetan, while BRDARTEN stands for Tibetan Basic Letters Plus BdaRten Coding Model in the diagrams below.

***Figure 3: BrdaRten Model for Migration, Processing and Rendering***

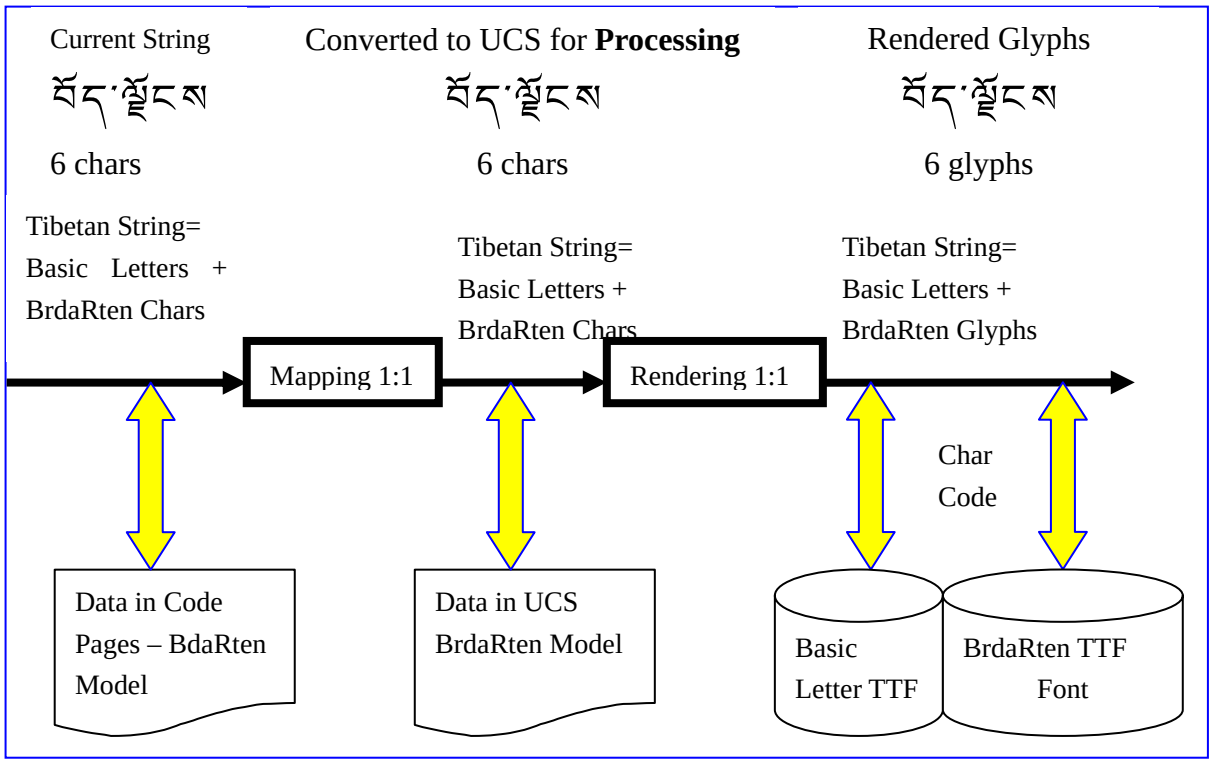
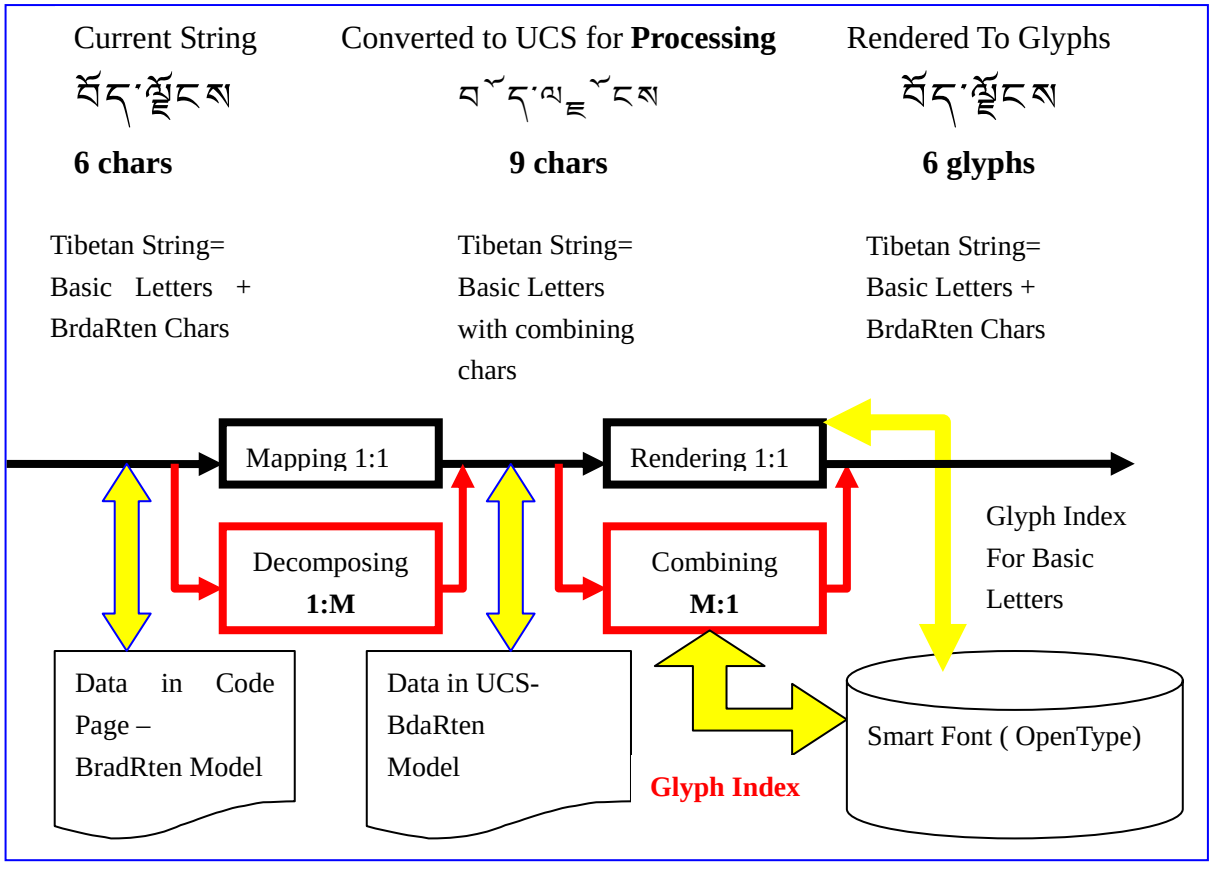



***Figure 4: Migration From BrdaRten Model to 0FXX Model for For Precessing & Rendering***



(1) Customarily, the form in pure 0FXX string with no combining (see the left Tibetan string below) is not readable by ordinary people, even the string occasionally appears. Unfortunately, it will inevitably occur providing the rendering engine or OpenType font is missing or disabled. This kind of pure **OFXX string may be easily Romanized but not easily recognized by ordinary Tibetan people**.

བོད་ལྗོངས་ ལུ་ས vs. བོད་སྐྱོངས་ ལུ་ས

(2) Given the fact that huge amount e-data with BrdaRten, it will be more costly for Chinese Tibetan to **migrate to UCS/**0FXX mode.

a) It is required to decompose all BrdaRten to be pure 0FXX string; **one to many mapping must be done**.

b) It causes Not-Equal-Length-Conversion as a result, **the DATA TYPE and Data Base Structure (say, the field length) would change accordingly**.

No one doubts the POSSIBILITY of such decomposing or mapping (as N2628 stated), but the COST and WHAT we get after the complex conversion?

(3) For character **processing**, apparently 0FXX makes the processing more complicated in cursor moving, character deleting, inserting or replacing. The complex-script rendering mechanism becomes required. While the BrdaRten model makes processing so simple as English or Chinese processing that all English and Chinese software could be used without major change. The reason is that **BrdaRten model changes Tibetan from so-called Complex-Script to be the Simple-Script in a sense**.

(4) For **storage**, BrdaRten takes less space than 0FXX string.

(5) For **IME**, BrdaRten is simple like Pinyin, and several implementations have enabled word-phrase-based input . While 0FXX needs multiple keyboard pages to manipulate that reduces user-friendliness.

(6) For **OCR**, in character segmentation stage, BdaRten needs vertical cutting only (see the left picture below), while the **0FXX needs both vertical and horizontal segmentation** ( see the right picture below ), makes implementation more difficult.

BrdaRten  0FXX 

(7) In **post OCR** processing, the 1:1 corresponding between the original image and the recognized characters would be helpful to **proof-reading** applications. The BrdaRten makes this easy to implement while **0FXX may lose the corresponding**.

(8) For **sorting**, 0FXX seems much simpler than BrdaRten, however, the collation of Tibetan has been developed in China.

(9) For searching, there is no apparent difference between 0FXX and BrdaRten. The example, search letter KA, in N2638 is not convincing since such searching is

semantically meaningless therefore the there is no need to decompose BrdaRten any more.

(10) Finally, for rendering, 0FXX needs Complex-Script Engine and “Advanced” OpenType, but BrdaRten does not.

**In summary, 0FXX increases the conversion cost to migrate legacy to UCS, and also rises the complexity in processing, storage, and rendering. In years, 0FXX seems not the most suitable coding model, at least for China. Chinese Tibetan’s choice is BrdaRten. As far as INTEROPERABILITY and INTERCHANGEABILITY, they are our goal. It is for INTEROPERABILITY with outside, we are proposing BrdaRten to be encoded in BMP not only in a National Standard, even within the UCS framework.**

==========

## The contributors of this paper are from the following companies, institutes, organizations and universities:

Translation & Editing Bureau, Tibet Autonomous Region, China
西藏自治区翻译局
Language Committee Office, Tibet Autonomous Region, China
西藏自治区藏语文工作委员会办公室
Tibet University
西藏大学
Nationalities Publishing House
民族出版社
Publishing House, China Tibetology Research Center
藏学研究中心出版社
Qinghai Normal University
青海师范大学
Beijing Founder Electronics Co., Ltd.
北京北大方正电子有限公司
Weifang Qingniao-Huaguang Co., Ltd.
北大青鸟-华光科技股份有限公司
UniHan Digital Technology Co., Ltd.
书同文数字化技术有限公司
Northwest Minorities University
西北民族大学
Software Institute, Chinese Academy of Sciences
中国科学院软件研究所
China Electronics Standardization Institute
中国电子标准化研究所

Standardization Institute of China
中国标准化研究院
Institute of Ethnology & Anthropology, Chinese Academy of Social Sciences
中国社会科学院 民族学与人类学研究所
Department of Language Information Administration, Ministry of Education, PRC
教育部语言文字信息管理司
Department of Culture & Publicity, State Ethnic Affairs Commission, PRC
国家民族事务委员会文化宣传司
Department of Electronics Production, Ministry of Information Industry, PRC
信息产业部 电子产品管理司